

On Discrete Infinite Horizon MPDs

- The solution to infinite horizon MPD problems is mathematically equivalent to computing a fixed point of the Bellman operator $V = \Gamma(V)$.
- This fixed point problem can also be presented as the problem of finding a zero to the non-linear functional $F(V) = 0$, where $F = [I - \Gamma]$.
- There are two main ways to compute fixed points to contraction mappings, successive approximations and the Newton-Kantorovich method. The latter is an application to non-linear equations of the familiar Newton method that uses Taylor series approximations around a current iteration to compute the zeros to the equations by minimizing a quadratic model.
- Most of the other methods that have been proposed to solve infinite horizon MPDs are applications of more general methods for solving systems of non-linear equations.
- Policy iteration is regarded as one of the fastest methods for computing V and the associated decision rule α . This is especially true when the number of states is large and the discount factor, β , is close to 1.
- As we mentioned we start by choosing an arbitrary initial policy, for example

$$\alpha_0(s) = \arg \max_{a \in A(s)} [u(s, a)]. \quad (1)$$

- We then carry out the policy valuation step to compute the value function V_{α_0} implied by the stationary decision rule. This requires solving the linear system

$$V_{\alpha_0} = u_{\alpha_0} + \beta M_{\alpha_0} V_{\alpha_0}, \quad (2)$$

where M_{α} is the Markov operator which in finite state problems reduces to a transition probability matrix with (i, j) element equal to

$$M_{\alpha}(i, j) = p(s' = s_j | s_i, \alpha(s_i)). \quad (3)$$

In maybe more familiar form the Markov operator is the linear conditional expectations operator

$$M_{\alpha}V(s) = \int V(s')p(ds' | s, \alpha(s)). \quad (4)$$

- Once we obtain the solution V_{α_0} a policy improvement step is used to generate an updated policy α_1 :

$$\alpha_1(s) = \arg \max_{a \in A(s)} \left[u(s, a) + \beta \sum_{s'=1}^S V_{\alpha_0}(s') p(s'|s, a) \right]. \quad (5)$$

- Given α_1 we would continue the cycle of policy valuation and policy improvement steps until the first iteration k such that $\alpha_k = \alpha_{k-1}$ or $V_{\alpha_k} = V_{\alpha_{k-1}}$. By the theory we have already studied regarding MDPs, the decision rule $\alpha = \alpha_k$ is optimal, and this method converges to the optimal rule in a finite number of iterations.
- As we mentioned the last time the reason for the fast convergence of this method is related to the rapid quadratic convergence rates of Newton's method for nonlinear equations. In fact it has been shown that policy iteration is a form of the Newton-Kantorovich method for finding a zero to the non-linear mapping $F : R^S \rightarrow R^S$ defined by $F(V) = [I - \Gamma](V)$.
- In many applications the number of policy iteration steps required to find the optimal policy appears to be independent of the number of states. However, the amount of work per iteration does depend on the number of states and is larger than for successive approximations, since we have to solve exactly the non-linear system. This is why parametric policy iteration has been suggested as a potentially faster extension.